The Effectiveness of Oversampling Low Income Households in the Survey of Income and Program Participation

Tiwanda M. Allen, Rita J. Petroni, Rajendra P. Singh
U.S. Bureau of the Census

To be presented at the American Statistical Association Meetings, August 1993.

I. INTRODUCTION

A. Background

The goal of the Survey of Income and Program Participation (SIPP) has been to provide policy makers with accurate and comprehensive information about the economic situation of persons and households in the noninstitutionalized U.S. population. Over the years, budget constraints dictated a reduction in the SIPP panel size. As data from the reduced panels became available, analysts found it more difficult to conduct meaningful analysis of government programs for the low income population. In response to analysts concerns about the diminished usefulness of the SIPP data to meet its goal, the Census Bureau pursued (1) various budget initiatives to increase the sample to its original size and (2) oversampling of the low income population. (King, 1990a.)

This paper describes the oversample design for the 1990 SIPP panel which the Census Bureau introduced in February through May 1990 and interviewed through June to September 1992. It examines the effectiveness of this oversample design in reaching the SIPP goal. Results of this paper can provide guidance to organizations considering to oversample.

B. Design of the 1990 SIPP Oversample Panel

The Census Bureau originally planned to introduce a 1990 SIPP panel of about 20,000 households selected with equal probability. Instead, the Bureau introduced a panel of 23,600 households which included an oversample of the low income population. Initially, we wanted to use income data to oversample the low income population. However, due to time constraints, this was operationally impossible. As a result, the Census Bureau used demographic characteristics of those who were occupying the sample housing units during February - May 1989 as auxiliary variables. These characteristics are: Black (BLK), Hispanic (HIS), and female headed with no spouse present living with relatives (FHNSP) households. Such households tend to have higher poverty rates than the general population. (King, 1990a.)

The 1990 oversample panel consists of the following three components:

Components of Oversample Panel	Number of Eligible Households
Households in addresses originally to be first interviewed in the 1990 panel.	19,700
Households associated with sample addresses which were to first be interviewed in February through May 1989 (i.e., households originally to be in the 1989 panel[1]) and were at that time headed by a Black, Hispanic, or FHNSP.	2,700
Households in one-ninth of all other 1989[1] panel sample addresses.	1,200

C. How Successful was the 1990 Oversample Design?

This paper examines the success of the Census Bureau's approach in increasing the number of low income cases and the impact of oversampling on the reliability of cross-sectional estimates at the beginning and end of the 1990 oversample panel. The oversample approach has been successful in increasing the number of low income cases. Black, Hispanic, and FHNSP headed households are good predictors of who will have low income at the beginning and end of a two and a half year panel. In addition, we found that addresses occupied by a Black, Hispanic, or FHNSP head in February through May 1989 tended also to be occupied by a Black, Hispanic, or FHNSP head in February 1990 through September 1992. The Census Bureau's oversample approach is also generally successful in increasing the reliability of low income-related and other SIPP estimates.

II. DATA ANALYSIS

A. Methodology

---

[1] The Census Bureau attempted to interview households in all sample addresses from the 1989 panel in February 1989 through January 1990. After January 1990, we did not interview for the 1989 panel. However, for the 1990 oversample panel, we interviewed the 1989 panel households included in the 1990 oversample panel.

In the following sections we analyze the stability of the characteristics of addresses and housing units with respect to auxiliary variables and low income status, increases in the number of low income cases, and the reliability of various characteristics after being in sample one year and then after an additional two and a half years. In the remainder of the paper we simply refer to "the characteristics of addresses and housing units with respect to auxiliary variables" as "auxiliary variables."

If our auxiliary variables are stable after one year, then we expect this to be the case after an additional two and a half years since we follow Wave 1 sample persons rather than addresses. Also, because of the way we oversampled, if our auxiliary variables are stable after one year, then the increase in low income cases should be approximately equal to the proportion of low income cases in the population times the number of cases from the 1989 panel.

One goal of the oversampling was to reduce the variances of low income-related estimates without a significant adverse affect on the variances of other SIPP estimates. Although our method increases the sample size for all population groups, the design introduces differential weights between cases from the original 1989 and 1990 panels. Since increased sample sizes decrease variances and differential weights increase them, we examine the actual variances to evaluate whether we met our goal.

In section B, we analyze changes in stability between Wave 1 of the 1989 panel and Wave 1 of the 1990 panel (Wave 1 is the interview months from February to May of 1989 and 1990, respectively); and between Wave 1 and Wave 8 of the 1990 panel (Wave 8 is the interview months from June to September of 1992). In section C, we examine increases in number of low income cases due to the 1989 panel cases at Waves 1 and 8 of the 1990 panel. In both sections, we examine unweighted counts.

Approximately, 3500 addresses from the 1989 panel were initially interviewed in Wave 1 of the 1990 panel. Due to unavailable data and sample loss, our analysis at different stages will have fewer than 3500 cases. It is necessary to obtain for each original 1989 panel address, household income and the race , sex, ethnicity, and marital status of the reference person. The availability of this information for each wave in our analysis, allows us to identify oversample addresses and analyze changes that are taking place for each address. However, as we match addresses between Wave 1 of the 1989 and Wave 1 of the 1990 panels, we loose addresses because not all of the necessary information is available for both.

In section D, we compare variances for the 1990 oversample design to variances when the 1989 panel cases are excluded. We computed the variances using the half sample replication option of VPLX. (Fay, 1990) For the 1990 oversample

design, we used our normal SIPP weighting procedures. (King, 1990 b and c). The weights include several adjustments to the baseweights. Two of these are: 1) an adjustment for combining samples from both the 1989 and 1990 panels; and 2) a raking ratio adjustment to account for population counts by age, race, sex, and household relationship. (King, 1990 b and c.) We derived weights for estimates which exclude the 1989 panel cases by:

- dividing out the combining and raking ratio adjustments for each original 1990 panel case,
- computing a new raking ratio adjustment, and
- multiplying by the new raking ratio adjustment.

B. Stability

Since the oversample cases from the 1989 panel were chosen based on our auxiliary variables, we will first analyze the stability of these variables.

Table 1 shows that 89% of the addresses that were BLK-HIS-FHNSP in Wave 1 of the 1989 panel were also in the same group in Wave 1 of the 1990 panel. Table 2 shows the stability of each of the variables separately. The BLK and HIS addresses are most stable with 94% and 87%, respectively, remaining in the same group. The "other" and FHNSP groups have about 70% of their addresses remaining in the same group in February to May 1990.

As mentioned earlier, if the auxiliary variables are stable after one year, we would expect a similar stability after an additional two and a half years in sample. Tables 3 and 4 show the stability of the variables from Wave 1 to Wave 8 of the 1990 panel. Since after the initial Wave 1 interview we follow persons instead of addresses, our results are as expected. Factors such as marriage, divorce, and death which affect the reference person for a household can account for changes in the auxiliary variables from Wave 1 to Wave 8.

From tables 1-4, we can calculate that over three and a half years about 81% of the households that were BLK-HIS-FHNSP headed in Wave 1 of 1989 were also BLK-HIS-FHNSP headed in Wave 8 of the 1990 panel. Of this group, FHNSP households are least stable. After three and a half years, only about 52% of such households are still classified as FHNSP.

In addition to the stability of the auxiliary variables we were also interested in the stability of income status using these variables after a year and then after an additional two and a half years. We classify a household into low income status if the household income is less than 125% of their poverty threshold.

From table 5 we calculated that 71% of BLK-HIS-FHNSP headed households that had low income status in Wave 1 of 1989 had the same status a year later. For the "other" households with low income status at Wave 1 of the 1989 panel, 41% of the households maintained the same status a year later. As for the households above 125% of their poverty thresholds in Wave 1 of 1989, about 12% of both the BLK-HIS-FHNSP and "other" households had low income status in Wave 1 of the 1990 panel.

Similar analysis was done for the two and a half year period from Wave 1 to Wave 8 of the 1990 panel. From table 6, 70% of BLK-HIS-FHNSP headed households with low income status in Wave 1 of 1990 had the same status in Wave 8 of the 1990 panel. This is about the same percentage as the one year analysis. The results were also similar between the one year and two and a half year analysis for the BLK-HIS-FHNSP headed households with incomes above 125% of the poverty threshold in Wave 1 of the 1990 panel, but have low income status in Wave 8 of the 1990 panel. After two and a half years, about 55% of both the "other" households that had and did not have low income status in Wave 1 of the 1990 panel had low income status in Wave 8 of the panel.

Over three and a half years, about 50% of the BLK-HIS-FHNSP households with low income status in Wave 1 of 1989, had the same status in Wave 8 of the 1990 panel.

The above analyses display the success and stability of our auxiliary variables and income status when using these variables. However, our original desire was to supplement our 1990 panel with only low income cases. From the data we had available we decided to simulate and analyze an oversample design based on income data. We would then compare the stability of the income status to that of the current oversample design.

In order to perform this additional analysis, we used data from Waves 1 and 8 of the 1990 panel. After removing the supplemented 1989 panel cases, we determined the income status of the original 1990 panel households. Table 7 shows the results of our analysis. This two and a half year analysis shows that 61% of the households with low income status in Wave 1 of 1990 had the same status in Wave 8 of the 1990 panel.

In constructing a similar table for our auxiliary variables for the supplemented 1989 panel cases, table 8, we find that 67% of the households with low income status in Wave 1 of 1990 had the same status in Wave 8. In addition, these cases have actually been in sample longer than two and a half years since they were originally interviewed in wave 1 of the 1989 panel.

Therefore, our current oversample design provided better results than the originally planned design.

C. Sample Size

From Table 5 we can calculate that 31% of the BLK-HIS-FHNSP headed households from the 1989 panel cases are considered low income households, while only 12% of the "other" households are low income. Combined, 24% of the cases taken from the 1989 panel are low income households in Wave 1 of the 1990 panel.

The BLK-HIS-FHNSP addresses from the 1989 Panel are providing a 44% increase in the number of low income households in Wave 1 of the 1990 panel while the "other" addresses are providing a 10% increase. Totally, the 1989 panel addresses have increased the number of low income cases 26% for Wave 1 of the 1990 panel. This was achieved by only increasing our sample size 17% for the original 1990 panel.

Similar results from the 1989 panel cases were obtained at Wave 8 of the 1990 panel.

D. Reliability

We analyzed two sets of approximately 1700 cross-sectional national estimates and variances. One set was produced using the oversample panel cases, while the other set does not include the 1989 panel cases (the non-oversampling panel). The sets of estimates available are for the first quarter of 1990. This allows us to evaluate the reliability of estimates at the beginning of the 1990 panel.

Overall, variances for 74% of the 1700 estimates from the oversample design are smaller than the variances for non-oversampling. The majority (66%) of the variances from the oversample design are at least 10% smaller than the variances from the non-oversampling design. The oversample approach has positively affected these estimates by decreasing their variances.

In addition to the above analysis, we wanted to compare the same variable characteristics for different populations such as the Total, Black, Hispanic, and persons aged 65 and over (65+), to see the affect the oversample approach is having on these different groups.

For example, we analyzed the variances that include the oversample and the variances that do not include the 1989 panel cases. Overall, we found that the variances for the oversample approach were smaller than the non-oversample approach's for the majority of the estimates analyzed for the Total, Black,

Hispanic, and 65+ populations. These estimates included low income and non-low income type estimates. Therefore, in general, the oversample approach is improving the variances of low income estimates, without adversely affecting the general SIPP estimates.

III. RESULTS AND CONCLUSIONS

To analyze the effectiveness of the 1990 panel's oversample design we studied the stability of the auxiliary variables, the increase in sample size, and the reliability of our SIPP estimates.

The results from the stability analysis showed that the characteristics of the occupants of sample households or addresses with respect to auxiliary variables are stable after one year and even more so after a two and half year period. This stability is mainly due to the type of variables selected and the fact that our survey follows Wave 1 persons not Wave 1 addresses. After three and a half years in sample, 81% of the households that were BLK-HIS-FHNSP headed in Wave 1 of the 1989 panel were BLK-HIS-FHNSP headed in Wave 8 of the 1990 Panel.

The 1989 panel addresses included in the 1990 panel increased the sample size about 17%. With this increase we were able to obtain approximately a 26% increase in the number of low income cases in both Wave 1 and Wave 8 of the 1990 Panel.

Generally the reliability of our 1990 first quarter low income type estimates have improved along with the reliability of our other 1990 first quarter SIPP estimates. Recall that 74% of the variances from the oversample design are smaller than the non-oversampling panel variances for the 1700 cross-sectional estimates.

Initially we wanted to use income data to identify low income households. However, due to time and budget constraints we were unable. Analyses in this paper showed that if we had used income data instead of the selected auxiliary variables our results would have been less successful at later waves.

Results of the research presented here show that the SIPP 1990 oversampling method was successful for the SIPP cross-sectional estimation purposes both at the beginning and end of the panel.

The oversampling approach was successful in increasing the number of cross-sectional low income cases and improving the reliability of cross-sectional low income estimates without a significant adverse affect on other cross-sectional national estimates. To complete the analysis, we should:

- compute and analyze variances for estimates obtained later in the panel,

- research the impact on longitudinal variances, and
- research the impact of oversampling on cross-sectional and longitudinal variances when the oversample design and non-oversample design are the same size.

These results also suggest that, at least when the goal is to oversample for low income households over a period of time, screening using income is not the best method. Using auxiliary variables for which the characteristics of housing units are stable over time and are correlated with low income is the better approach in this case. Research is needed to determine which is the better method when the goal is to oversample low income households for a one time survey that is to be carried out close to the time of screening.

REFERENCES

Fay, R. (1990), "VPLX: Variance Estimates for Complex Samples," 1990 Proceedings of the Survey Research Methods Section of the American Statistical Association.

King, K. (1990a), "SIPP: Restructuring the 1989 and 1990 Panels," Internal Census Bureau Memorandum for Documentation, January 31, 1990.

King, K. (1990b), "SIPP 90: Cross-Sectional Weighting Specifications for Wave 1," Internal Census Bureau Memorandum for Courtland from Waite, September 26, 1990.

King, K. (1990c), "SIPP 90: Cross-Sectional Weighting Specifications for Second and Subsequent Waves," Internal Census Bureau Memorandum for Courtland from Waite, November 5, 1990.

Table 1: Cross-Tabulation of the Type of Address in Wave 1 of the 1989 Panel Versus Wave 1 of the 1990 Panel.

1989 Panel, Wave 1	1990 Panel, Wave 1 Type of Address		
Type of Address	BLK-HIS-FHNSP	Other	Total
BLK-HIS-FHNSP	1458 89.17	177 10.83	1635
Other	418 29.33	1007 70.67	1425
Total	1876	1184	3060

Table 2: Cross-Tabulation of the Type of Address in Wave 1 of the 1989 Panel Versus Wave 1 of the 1990 Panel by Group.

1989 Panel, Wave 1	1990 Panel, Wave 1 Type of Address				
Type of Address	Black	FHNSP	Hispanic	Other	Total
Black	749 94.45	4 0.50	6 0.76	34 4.29	793
FHNSP	6 1.43	296 70.31	8 1.90	111 26.37	421
Hispanic	14 3.33	7 1.66	368 87.41	32 7.60	421
Other	100 6.97	216 15.20	102 7.18	1007 70.65	1425
Total	869	523	484	1184	3060

Table 3: Cross-Tabulation of the Type of Address in Wave 1 of the 1990 Panel Versus Wave 8 of the 1990 Panel.

1990 Panel, Wave 1	1990 Panel, Wave 8 Type of Address		
Type of Address	BLK-HIS-FHNSP	Other	Total
BLK-HIS-FHNSP	1660 91.81	148 8.19	1808
Other	47 3.86	1172 96.14	1219
Total	1707	1320	3027

Table 4: Cross-Tabulation of the Type of Address in Wave 1 of the 1990 Panel Versus Wave 8 of the 1990 Panel by Group.

1990 Panel, Wave 1	1990 Panel, Wave 8 Type of Address				
Type of Address	Black	FHNSP	Hispanic	Other	Total
Black	803 99.26	0 0.00	1 0.12	5 0.62	809
FHNSP	2 0.38	391 73.91	0 0.00	136 25.71	529
Hispanic	0 0.00	1 0.21	462 98.30	7 1.49	470
Other	3 0.25	43 3.53	1 0.08	1172 96.14	1219
Total	808	435	464	1320	3027

Table 5: Income Levels for Auxiliary Variables in Wave 1 of the 1989 Panel Versus Wave 1 of the 1990 Panel.

1989 Panel, Wave 1		Income Level: 1990 Panel, Wave 1				
		Below 125%		Above 125%		
Income Level	Type of Address	BLK-HIS-FHNSP	Other	BLK-HIS-FHNSP	Other	Total
Below 125%	BLK-HIS-FHNSP	360	11	117	33	521
	Other	29	45	36	70	180
Above 125%	BLK-HIS-FHNSP	124	8	857	125	1114
	Other	74	82	279	810	1245
Total		587	146	1289	1038	3060

Table 6: Income Levels for Auxiliary Variables in Wave 1 of the 1990 Panel Versus Wave 8 of the 1990 Panel.

1990 Panel, Wave 1		Income Level: 1990 Panel, Wave 8				
		Below 125%		Above 125%		
Income Level	Type of Address	BLK-HIS-FHNSP	Other	BLK-HIS-FHNSP	Other	Total
Below 125%	BLK-HIS-FHNSP	365	18	143	19	545
	Other	9	73	5	60	147
Above 125%	BLK-HIS-FHNSP	147	13	1005	98	1263
	Other	9	65	24	974	1072
Total		530	169	1177	1151	3027

Table 7: Cross-Tabulation of Income Level of Addresses in Wave 1 of the 1990 Panel Versus Wave 8 of the 1990 Panel, without the 1989 Panel Cases.

1990 Panel, Wave 1	1990 Panel, Wave 8 Income Level		
Income Level	Below 125%	Above 125%	Total
Below 125%	1442 60.95	924 39.05	2366
Above 125%	1032 7.61	12531 92.39	13563
Total	2474	13455	15929

Table 8: Cross-Tabulation of Income Level of Auxiliary Variables from the 1989 Panel in Wave 1 of 1989 Versus Wave 8 of the 1990 Panel.

1990 Panel, Wave 1	1990 Panel, Wave 8 Income Level		
Income Level	Below 125%	Above 125%	Total
Below 125%	465 67.20	227 32.80	692
Above 125%	234 10.02	2101 89.98	2335
Total	699	2328	3027